Emily Alsentzer & Julie Zhu
Project Proposal
Stats 205

The Epidemic of Police Killings

Introduction

In 2014, a series of highly publicized police killings of unarmed Black men prompted national discourse on racial biases of policing in the United States. The killing of Michael Brown sparked protests and riots in Ferguson, Missouri[1], and the chokehold death of Eric Garner[2], the drive-by police shooting of twelve-year-old Tamir Rice[3], and the crippling condition of Freddie Gray[4] have repeatedly brought questions of police brutality and racialized violence to the forefront. After Trayvon Martin's killer was acquitted, Alicia Garza, Patrisse Cullors, and Opal Tometi founded the BlackLivesMatter movement, which they describe as "an ideological and political intervention in a world where Black lives are systematically and intentionally targeted for demise[5]." The #BlackLivesMatter movement continues to focus public attention on police brutality directed against Blacks, and it has influenced the discourse of the 2016 presidential election.

In light of these recent events, we hope to use police killing data collected by *The Guardian* as well as federal Census data in order to better understand the nature of police killings in the United States. In particular, we hope to elucidate what socioeconomic, environmental, political, and racial factors influence police killings, and we plan to examine the factors affecting the media attention of the police killings from the past two years. Specifically, we plan to address the following questions:

1. Are there racial disparities in the number of people killed by police?
2. Are there differences in the circumstances of the police killing among races?
3. Which socioeconomic, environmental, or political factors influence the number of police killings in a region?
4. How do race, location, threat level, and presence of mental illness affect media attention of the police killing?

Data

The discourse surrounding these events has raised awareness of the lack of data on police killings. Although the number of police officers killed in the line of duty is well documented[7], there is limited reliable information about the number of individuals killed by police[8]. In the wake of Ferguson, *The Guardian* newspaper launched an effort called "The Counted" in June 2015 to compile information about police killings[8]. Crowd sourced platforms such as "Fatal Encounters"[9] and "Killed by Police" have existed since 2012, but "The Counted" represents the

first effort to combine professional reporting with reader and user participation (of course, with verified information only).

We will use data[10] from "The Counted," a similar database by the Washington Post, and Census data[11] in order to conduct our analyses. "The Counted" and the Washington Post database both include deaths arising from encounters from law enforcement. However, "The Counted" includes those who died in custody whereas deaths of people in custody, fatal shootings by off-duty officers, and non-shooting deaths were not tracked in the Washington Post database[12]. Furthermore, self-inflicted deaths during encounters with the police are not counted in this data set, and mass shootout incidents such as the one in Waco, Texas in 2015 were also excluded. The name, age, gender, race/ethnicity, day of killing, street address of killing, manner of death, law enforcement agency, and whether the individual was armed or not were recorded in the data. It is important to note that that ethnicity and whether the individual was armed is self-reported by the law enforcement agency, which means that it could not be accurate.

The Washington Post data includes information on the specifics of killings from 2015-2016 including name, date, manner of death, city, state, and demographic data such as age, race, and gender. The database also records whether the individual was armed or fleeing, the threat level of the individual, whether there were signs of mental illness, and whether the police were wearing body cameras. Although the database was curated by journalists, the information is not complete nor is it completely accurate. Of the 1,305 total police killings from 2015-2016, there are 64 individuals with missing race information, 21 with missing ages, and 13 with missing information about whether the individual was fleeing. There is also no public information how signs of mental illness or ethnicity was determined.

Lastly, The American Community Survey (ACS) is a survey conducted by the U.S. Census Bureau every year. Addresses from every state, the District of Columbia, and Puerto Rico are sampled. The survey includes information on demographic, economic, social, and housing measures, including employment status, health insurance coverage, educational attainment, and rent statistics[13]. For this project, we will use poverty rates, educational attainment, and other socioeconomic factors in order to assess what contributes to increased police killings. The survey had a response rate of approximately 96% in 2014, and approximately 98.4% of all housing units or 92.5% of the entire population were covered in the survey[14].

Methods

Our data largely consists of categorical data. The only non-categorical data found in the dataset is the age of the deceased, which has a non-normal distribution (Shapiro-Wilk test p-value = 2.2e-16, Figure 1). Given the nature of our dataset, nonparametric statistical tests will likely be the most useful. Nonparametric tests are advantageous in that they do not rely on assumptions about the shape of the probability distribution of the data, and they handle outliers

well[15]. If the data do not constitute a random sample from a larger population, as in the case of the police killing data, then standard parametric techniques may no longer be appropriate[16]. Although nonparametric methods can be useful, they may lack power compared to parametric approaches, which can be problematic with small sample sizes or if normality assumptions hold. Additionally, nonparametric methods are designed more for hypothesis testing that estimation, which means that additional work, such as using the bootstrap, is needed to generate confidence intervals[15]. Furthermore, non-parametric methods generally work better with larger data sets in order to reduce noise the variance of our estimates. However, given the categorical nature and size (around 1,000 rows) of our data, we believe that non-parametric methods will prove most useful.

Before performing analyses on our specific questions, we will first perform a multiple correspondence analysis in order to better understand the data. We will use the built-in R package, factoMineR, which will allow us to identify groups of individuals and examine relationships between variables. We can then proceed with further analyses. Following is a list of our research questions, from least to most complex, with proposed methods to address each question.

1. Are there racial disparities in the number of people killed by police?

In order to address this question, we will perform a chi-squared test, similar to the unfair dice roll example in lecture, in order to ascertain whether the proportions of people killed per race is different than what we would expect according to the general population. We will use bootstrapping in order to generate a confidence interval.

2. Are there differences in the circumstances of the police killing among races?

First, we will divide our data into groups by race. Then, we will use a proportion test to compare the percent of individuals with signs of mental illness, unarmed, and fleeing across different races and ethnicities. Because we will be performing multiple comparisons, we will use Bonferroni correction to adjust the significance threshold.

3. Which socioeconomic, environmental, or political factors influence the number of police killings in a region?

By joining the Census data with the police killings data, we can link information such as educational attainment and poverty spatially with the police killings. This enables us to use the Census data to create a model of the number of police killings in an area for inference. In order to have enough data for this analysis, we will join the datasets from the Guardian and the Washington Post. First, we must determine an appropriate spatial scale for this analysis after analyzing the data. Then, we will perform nonlinear regression in order to fit a model to our data and use feature selection in order to avoid model overfitting.

4. How do race, location, threat level, and presence of mental illness affect media attention of the police killing?

To begin, we will scrape tweets from Twitter using an open source python script[17]. Using these methods, we can specify a start and end date for the range of tweets. We will examine tweets that mention the name of the deceased beginning at the day of death and ending six months later, and the tweets will be filtered according to inclusion and exclusion word criteria. This will enable us to quantify the social media discussion surrounding each police killing by number of tweets. Because our estimate of the number of tweets for each individual may be limited by the specificity of our inclusion and exclusion criteria, we will categorize each individual into low, moderate, or high media attention depending on the number of tweets. We will then use log-linear analysis, which uses backwards feature selection to construct a model such that our data is accounted for by a minimum number of variables. This will allow us to determine which factors most explain heightened media attention. Finally, with time permitting, we hope to perform sentiment analysis to determine which factors may contribute to different types of responses over Twitter.

Timeline

Below is our timeline for the rest of the quarter. We plan to parallelize work in order to be most efficient. If time permits, we hope to conduct sentiment analysis on the tweets following the police killings. We would also like to either quantify discrepancies between the different police killing databases or use mapping software to create a heat map of the US of police killing factors using multivariate spatial nonparametric modeling using kernel processes mixing[18].

Week in Quarter	**Task**
Week 6	Data cleaning, Merge Police Killings and ACS Databases
Week 7	Multiple Correspondence Analysis, Questions 1 and 2
Week 8	Extract Tweets, Question 3
Week 9	Sentiment Analysis, Question 4
Week 10	Final Report

Conclusion

By incorporating the police killings data with the Census data, we hope to investigate the underlying factors that contribute to police killings. Furthermore, by examining the media attention associated with police killings, we hope to not only characterize which types of police killings are likely to receive media attention, but also uncover which police killing scenarios may be less underrepresented in the media.

References

1. Buchanan, Larry, and Ford Fessenden. "What Happened in Ferguson?" *The New York Times*. The New York Times, 12 Aug. 2014. Web. 26 Apr. 2016. <http://www.nytimes.com/interactive/2014/08/13/us/ferguson-missouri-town-under-siege-after-police-shooting.html?_r=0>.
2. Goldstein, Joseph, and Nate Schweber. "Man's Death After Chokehold Raises Old Issue for the Police." *The New York Times*. The New York Times, 18 July 2014. Web. 26 Apr. 2016. <http://www.nytimes.com/2014/07/19/nyregion/staten-island-man-dies-after-he-is-put-in-chokehold-during-arrest.html?smid=pl-share>.
3. McCarthy, Tom. "Tamir Rice: Video Shows Boy, 12, Shot 'seconds' after Police Confronted Child." *The Guardian*. Guardian News and Media, 26 Nov. 2014. Web. 26 Apr. 2016. <http://www.theguardian.com/us-news/2014/nov/26/tamir-rice-video-shows-boy-shot-police-cleveland>.
4. Yan, Holly, Carolyn Sung, Ashley Fantz, Kimberly Hutcherson, Eliott C. McLaughlin, Catherine E. Shoichet, Kevin Conlon, and Dana Ford24. "Police: Freddie Gray Didn't Get Timely Medical Care after Arrest." *CNN*. Cable News Network, 24 Apr. 2015. Web. 26 Apr. 2016. <http://www.cnn.com/2015/04/24/us/baltimore-freddie-gray-death/index.html>.
5. Garza, Alecia. "A Herstory of the #BlackLivesMatter Movement." *The Feminist Wire*. 07 Oct. 2014. Web. 26 Apr. 2016. <http://www.thefeministwire.com/2014/10/blacklivesmatter-2/>.
6. Gates, Kelly. "What the Absence of Data on Police Killings Revealed." Democracy, Citizenship and Constitutionalism Workshop. University of Pennsylvania, Philadelphia. 10 Dec. 2015. *Sas.upenn.edu*. Web. 24 Apr. 2016. <https://www.sas.upenn.edu/dcc/sites/www.sas.upenn.edu.dcc/files/GatesAbsenceOfDataPAPER.pdf>.
7. Tarm, Michael. "Are More Police Getting Killed? A Look at Officer Deaths." *The Seattle Times*. Frank A. Blethen, 02 Sept. 2015. Web. 26 Apr. 2016. <http://www.seattletimes.com/nation-world/a-look-at-recent-law-enforcement-deaths-across-the-u-s/>.
8. Swaine, Jon, Oliver Laughland, Jamiles Lartey, Kenan Davis, Rich Harris, Nadja Popovich, Kenton Powell, and Guardian US Interactive Team. "The Counted: People Killed by Police in the United States – Interactive." *The Guardian*. Guardian News and Media, 01 June 2015. Web. 26 Apr. 2016. <http://www.theguardian.com/us-news/ng-interactive/2015/jun/01/the-counted-police-killings-us-database>.
9. Burghart, D. Brian. Web log post. *Fatal Encounters*. 20 Oct. 2015. Web. 26 Apr. 2016.

10. "Washington Post Data Police Shootings." *GitHub*. 1 Apr. 2016. Web. 26 Apr. 2016. <https://github.com/washingtonpost/data-police-shootings>.
11. "American Community Survey (ACS)." *American Community Survey (ACS)*. U.S. Department of Commerce. Web. 26 Apr. 2016. <https://www.census.gov/programs-surveys/acs/>.
12. Tate, Julie, Jennifer Jenkins, Steven Richardson, John Muyskens, Kennedy Elliott, and Ted Mellnik. "How The Washington Post Is Examining Police Shootings in the U.S." *Washington Post*. The Washington Post, 1 Apr. 2016. Web. 26 Apr. 2016. <https://www.washingtonpost.com/investigations/how-the-washington-post-is-examining-police-shootings-in-the-us/2016/03/31/4d37f322-f5bf-11e5-8b23-538270a1ca31_story.html>.
13. "Subjects Included in the Survey." *American Community Survey*. U.S. Department of Commerce. Web. 26 Apr. 2016. <https://www.census.gov/programs-surveys/acs/guidance/subjects.html>.
14. "Response Rates." *American Community Survey*. U.S. Department of Commerce. Web. 26 Apr. 2016. <https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/response-rates/>.
15. Whitley, Elise, and Jonathan Ball. "Statistics Review 6: Nonparametric Methods." *Critical Care*. BioMed Central, 13 Sept. 2002. Web. 26 Apr. 2016. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC153434/>.
16. Dallal, Gerard E., PhD. "Nonparametric Statistics." *Nonparametric Statistics*. 23 May 2012. Web. 29 Apr. 2016. <http://www.jerrydallal.com/lhsp/npar.htm>.
17. Henrique, Jefferson. "Get Old Tweets." *GitHub*. 1 Feb. 2015. Web. 29 Apr. 2016. <https://github.com/Jefferson-Henrique/GetOldTweets-python>.
18. Fuentes, Michael, and Barton Reich. "Result Filters." *National Center for Biotechnology Information*. U.S. National Library of Medicine, 23 Jan. 2013. Web. 26 Apr. 2016. <http://www.ncbi.nlm.nih.gov/pubmed/24347994>.